Fair Regression with Wasserstein Barycenters

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil

Fair Regression with Wasserstein Barycenters

Evgenii Chzhen[1], **Christophe Denis**[2], **Mohamed Hebiri**[2]
Luca Oneto[3], **and Massimiliano Pontil**[4,5]
[1]LMO, Université Paris-Saclay, CNRS, Inria, [2]LAMA, Université Gustave Eiffel,
[3]DIBRIS, University of Genoa, [4]Istituto Italiano di Tecnologia, [5]University College London

Abstract

We study the problem of learning a real-valued function that satisfies the Demographic Parity constraint. It demands the distribution of the predicted output to be independent of the sensitive attribute. We consider the case that the sensitive attribute is available for prediction. We establish a connection between fair regression and optimal transport theory, based on which we derive a close form expression for the optimal fair predictor. Specifically, we show that the distribution of this optimum is the Wasserstein barycenter of the distributions induced by the standard regression function on the sensitive groups. This result offers an intuitive interpretation of the optimal fair prediction and suggests a simple post-processing algorithm to achieve fairness. We establish risk and distribution-free fairness guarantees for this procedure. Numerical experiments indicate that our method is very effective in learning fair models, with a relative increase in error rate that is inferior to the relative gain in fairness.

1 Introduction

A central goal of algorithmic fairness is to ensure that sensitive information does not "unfairly" influence the outcomes of learning algorithms. For example, if we wish to predict the salary of an applicant or the grade of a university student, we would like the algorithm to not unfairly use additional sensitive information such as gender or race. Since today's real-life datasets often contain discriminatory bias, standard machine learning methods behave unfairly. Therefore, a substantial effort is being devoted in the field to designing methods that satisfy "fairness" requirements, while still optimizing prediction performance, see for example [5, 10, 13, 16, 18, 21, 23, 25, 26, 28, 32, 45–47, 49] and references therein.

In this paper we study the problem of learning a real-valued regression function which among those complying with the Demographic Parity fairness constraint, minimizes the mean squared error. Demographic Parity requires the probability distribution of the predicted output to be independent of the sensitive attribute and has been used extensively in the literature, both in the context of classification and regression [1, 12, 20, 24, 34]. In this paper we consider the case that the sensitive attribute is available for prediction. Our principal result is to show that the distribution of the optimal fair predictor is the solution of a Wasserstein barycenter problem between the distributions induced by the unfair regression function on the sensitive groups. This result builds a bridge between fair regression and optimal transport, [see *e.g.,* 38, 41].

We illustrate our result with an example. Assume that X represents a candidate's skills, S is a binary attribute representing two groups of the population (*e.g.,* majority or minority), and Y is the current market salary. Let $f^*(x, s) = \mathbb{E}[Y|X{=}x, S{=}s]$ be the regression function, that is, the optimal prediction of the salary currently in the market for candidate (x, s). Due to bias present in the underlying data distribution, the induced distribution of market salary predicted by f^* varies across

Preprint. Under review.

the two groups. We show that the optimal fair prediction g^* transforms the regression function f^* as

$$g^*(x,s) = p_s f^*(x,s) + (1-p_s)t^*(x,s) \ ,$$

where p_s is the frequency of group s and the correction $t^*(x,s)$ is determined so that the *ranking* of $f^*(x,s)$ relative to the distribution of $X|S=s$ for group s (*e.g.,* minority) is the same as the ranking of $t^*(x,s)$ relative to the distribution of the group $s' \neq s$ (*e.g.,* majority). We elaborate on this example after Theorem 2.3 and in Figure 1. The above expression of the optimal fair predictor naturally suggests a simple post-processing estimation procedure, where we first estimate f^* and then transform it to get an estimator of g^*. Importantly, the transformation step involves only unlabeled data since it requires estimation of cumulative distribution functions.

Contributions and organization. In summary we make the following contributions. First, in Section 2 we derive the expression for the optimal function which minimizes the squared risk under Demographic Parity constraints (Theorem 2.3). This result establishes a connection between fair regression and the problem of Wasserstein barycenters, which allows to develop an intuitive interpretation of the optimal fair predictor. Second, based on the above result, in Section 3 we propose a post-processing procedure that can be applied on top of any off-the-shelf estimator for the regression function, in order to transform it into a fair one. Third, in Section 4 we show that this post-processing procedure yields a fair prediction independently from the base estimator and the underlying distribution (Proposition 4.1). Moreover, finite sample risk guarantees are derived under additional assumptions on the data distribution provided that the base estimator is accurate (Theorem 4.4). Finally, Section 5 presents a numerical comparison of the proposed method *w.r.t.* the state-of-the-art.

Related work. Unlike the case of fair classification, fair regression has received limited attention to date; we are only aware of few works on this topic that are supported by learning bounds or consistency results for the proposed estimator [1, 34]. Connections between algorithmic fairness and Optimal Transport, and in particular the problem of Wasserstein barycenters, has been studied in [12, 20, 24, 43] but mainly in the context of classification. These works are distinct from ours, in that they do not show the link between the optimal fair regression function and Wasserstein barycenters. Moreover, learning bounds are not addressed therein. Our distribution-free fairness guarantees share similarities with contributions on prediction sets [30, 31] and conformal prediction literature [42, 48] as they also rely on results on rank statistics. Meanwhile, the risk guarantee that we derive, combines deviation results on Wasserstein distances in one dimension [7] with peeling ideas developed in [3], and classical theory of rank statistics [40].

Notation. For any positive integer $N \in \mathbb{N}$ we denote by $[N]$ the set $\{1,\dots,N\}$. For $a,b \in \mathbb{R}$ we denote by $a \wedge b$ (*resp.* $a \vee b$) the minimum (*resp.* the maximum) between a and b. For two positive real sequences a_n, b_n we write $a_n \lesssim b_n$ to indicate that there exists a constant c such that $a_n \leq c b_n$ for all n. For a finite set $\mathcal{S}$ we denote by $|\mathcal{S}|$ its cardinality. The symbols $\mathbf{E}$ and $\mathbf{P}$ stand for generic expectation and probability. For any univariate probability measure μ, we denote by F_μ its Cumulative Distribution Function (CDF) and by $Q_\mu : [0,1] \to \mathbb{R}$ its quantile function (*a.k.a.* generalized inverse of F_μ) defined for all $t \in (0,1]$ as $Q_\mu(t) = \inf\{y \in \mathbb{R} \ : \ F_\mu(y) \geq t\}$ with $Q_\mu(0) = Q_\mu(0+)$. For a measurable set $A \subset \mathbb{R}$ we denote by $U(A)$ the uniform distribution on A.

2 The problem

In this section we introduce the fair regression problem and present our derivation for the optimal fair regression function alongside its connection to Wasserstein barycenter problem. We consider the general regression model

$$Y = f^*(X,S) + \xi \ , \tag{1}$$

where $\xi \in \mathbb{R}$ is a centered random variable, $(X,S) \sim \mathbb{P}_{X,S}$ on $\mathbb{R}^d \times \mathcal{S}$, with $|\mathcal{S}| < \infty$, and $f^* : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ is the regression function minimizing the squared risk. Let $\mathbb{P}$ be the joint distribution of (X,S,Y). For any prediction rule $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, we denote by $\nu_{f|s}$ the distribution of $f(X,S)|S=s$, that is, the Cumulative Distribution Function (CDF) of $\nu_{f|s}$ is given by

$$F_{\nu_{f|s}}(t) = \mathbb{P}(f(X,S) \leq t | S = s) \ , \tag{2}$$

to shorten the notation we will write $F_{f|s}$ and $Q_{f|s}$ instead of $F_{\nu_{f|s}}$ and $Q_{\nu_{f|s}}$ respectively.

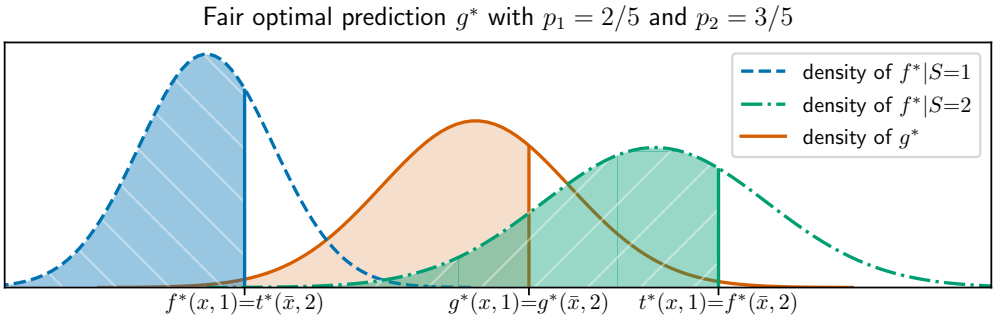

Figure 1: For a new point $(x, 1)$, the value $t^*(x, 1)$ is chosen such that the shaded `Green Area (//)` $= \mathbb{P}(f^*(X, S) \leq t^*(x, 1)|S = 2)$ equals to the shaded `Blue Area (\\)` $= \mathbb{P}(f^*(X, S) \leq f^*(x, 1)|S = 1)$. The final prediction $g^*(x, 1)$ is a convex combination of $f^*(x, 1)$ and $t^*(x, 1)$. The same is done for $(\bar{x}, 2)$.

Definition 2.1 (Wasserstein-2 distance)**.** *Let μ and ν be two univariate probability measures. The squared Wasserstein-2 distance between μ and ν is defined as*

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu,\nu}} \int |x - y|^2 \, d\gamma(x, y) \ ,$$

where $\Gamma_{\mu,\nu}$ is the set of distributions (couplings) on $\mathbb{R} \times \mathbb{R}$ such that for all $\gamma \in \Gamma_{\mu,\nu}$ and all measurable sets $A, B \subset \mathbb{R}$ it holds that $\gamma(A \times \mathbb{R}) = \mu(A)$ and $\gamma(\mathbb{R} \times B) = \nu(B)$.

In this work we use the following definition of (strong) Demographic Parity, which was previously used in the context of regression by [1, 12, 24].

Definition 2.2 (Demographic Parity)**.** *A prediction (possibly randomized) $g : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ is* fair *if, for every $s, s' \in \mathcal{S}$*

$$\sup_{t \in \mathbb{R}} \left| \mathbf{P}(g(X, S) \leq t | S = s) - \mathbf{P}(g(X, S) \leq t | S = s') \right| = 0 \ .$$

Demographic Parity requires the Kolmogorov-Smirnov distance between $\nu_{g|s}$ and $\nu_{g|s'}$ to vanish for all s, s'. Thus, if g is fair, $\nu_{g|s}$ does not depend on s and to simplify the notation we will write ν_g.

Recall the model in Eq. (1). Since the noise has zero mean, the minimization of $\mathbb{E}(Y - g(X, S))^2$ over g is equivalent to the minimization of $\mathbb{E}(f^*(X, S) - g(X, S))^2$ over g. The next theorem shows that the optimal fair predictor for an input (x, s) is obtained by a nonlinear transformation of the vector $(f^*(x, s))_{s=1}^{|\mathcal{S}|}$ that is linked to a Wasserstein barycenter problem [2].

Theorem 2.3 (Characterization of fair optimal prediction)**.** *Assume, for each $s \in \mathcal{S}$, that the univariate measure $\nu_{f^*|s}$ has a density and let $p_s = \mathbb{P}(S = s)$. Then,*

$$\min_{g \text{ is fair}} \mathbb{E}(f^*(X, S) - g(X, S))^2 = \min_{\nu} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \ .$$

Moreover, if g^* and ν^* solve the l.h.s. and the r.h.s. problems respectively, then $\nu^* = \nu_{g^*}$ and

$$g^*(x, s) = \left(\sum_{s' \in \mathcal{S}} p_{s'} Q_{f^*|s'} \right) \circ F_{f^*|s} \left(f^*(x, s) \right) \ . \tag{3}$$

The proof of Theorem 2.3 relies on the classical characterization of optimal coupling in one dimension (stated in Theorem A.1 in the appendix) of the Wasserstein-2 distance. We show that a minimizer g^* of the L_2-risk can be used to construct ν^* and vice-versa, given ν^*, we leverage a well-known expression for one dimensional Wasserstein barycenter (see *e.g.,* [2, Section 6.1] and Lemma A.2 in the appendix) and construct g^*.

The case of binary protected attribute. Let us unpack Eq. (3) in the case that $\mathcal{S} = \{1, 2\}$, assuming *w.l.o.g.* that $p_2 \geq p_1$. Theorem 2.3 states that the fair optimal prediction g^* is defined for all individuals $x \in \mathbb{R}^d$ in the first group as

$$g^*(x, 1) = p_1 f^*(x, 1) + p_2 t^*(x, 1), \text{ with } t^*(x, 1) = \inf \left\{ t \in \mathbb{R} \ : \ F_{f^*|2}(t) \geq F_{f^*|1}(f^*(x, 1)) \right\} \ ,$$

and likewise for the second group. This form of the optimal fair predictor, and more generally Eq. (3), allows us to understand the decision made by g^* at individual level. If we interpret (x, s) as the candidate's CV and candidate's group respectively, and $f^*(x, s)$ as the current market salary (which might be discriminatory), then the fair optimal salary $g^*(x, s)$ is a convex combination of the market salary $f^*(x, s)$ and the adjusted salary $t^*(x, s)$, which is computed as follows. ebgin If say $s{=}1$, we first compute the fraction of individuals from the first group whose market salary is at most $f^*(x, 1)$, that is, we compute $\mathbb{P}(f^*(X, S) \leq f^*(x, 1)|S{=}1)$. Then, we find a candidate $\bar{x}$ in group 2, such that the fraction of individuals from the second group whose market salary is at most $f^*(\bar{x}, 2)$ is the same, that is, $\bar{x}$ is chosen to satisfy $\mathbb{P}(f^*(X, S) \leq f^*(\bar{x}, 2)|S{=}2) = \mathbb{P}(f^*(X, S) \leq f^*(x, 1)|S{=}1)$. Finally, the market salary of $\bar{x}$ is exactly the adjustment for x, that is, $t^*(x, 1) = f^*(\bar{x}, 2)$. This idea is illustrated in Figure 1 and leads to the following philosophy: if candidates $(x, 1)$ and $(\bar{x}, 2)$ share the same group-wise market salary ranking, then they should receive the same salary determined by the fair prediction $g^*(x, 1) = g^*(\bar{x}, 2) = p_1 f^*(x, 1) + p_2 f^*(\bar{x}, 2)$. At last, note that Eq. (3) allows to understand the (potential) amount of extra money that we need to pay in order to satisfy fairness. While the unfair decision made with f^* costs $f^*(x, 1){+}f^*(\bar{x}, 2)$ for the salary of $(x, 1)$ and $(\bar{x}, 2)$, the fair decision g^* costs $2(p_1 f^*(x, 1){+}p_2 f^*(\bar{x}, 2))$. Thus, the extra (signed) salary that we pay is $\Delta = (p_2{-}p_1)(f^*(\bar{x}, 2){-}f^*(\bar{x}, 1))$. Since, $p_2{\geq}p_1$, Δ will be positive whenever the candidate $\bar{x}$ from the majority group gets higher salary according to f^*, and negative otherwise. We believe that the expression Eq. (3) could be the starting point for further more applied work on algorithmic fairness.

3 General form of the estimator

In this section we propose an estimator of the optimal fair predictor g^* that relies on the plug-in principle. The expression (3) of g^* suggests that we only need estimators for the regression function f^*, the proportions p_s, as well as the CDF $F_{f^*|s}$ and the quantile function $Q_{f^*|s}$, for all $s \in \mathcal{S}$. While the estimation of f^* needs labeled data, all the other quantities rely only on $\mathbb{P}_S$, $\mathbb{P}_{X|S}$ and f^*, therefore *unlabeled* data with an estimator of f^* suffices. Thus, given a base estimator of f^*, our post-processing algorithm will require only unlabeled data.

For each $s \in \mathcal{S}$ let $\mathcal{U}^s{=}\{X_i^s\}_{i=1}^{N_s} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X|S=s}$ be a group-wise unlabeled sample. In the following for simplicity we assume that N_s are *even* for all $s \in \mathcal{S}$[1]. Let $\mathcal{I}_0^s, \mathcal{I}_1^s \subset [N_s]$ be any fixed partition of $[N_s]$ such that $|\mathcal{I}_0^s|{=}|\mathcal{I}_1^s|{=}N_s/2$ and $\mathcal{I}_0^s{\cup}\mathcal{I}_1^s{=}[N_s]$. For each $j \in \{0, 1\}$ we let $\mathcal{U}_j^s{=}\left\{X_i^s \in \mathcal{U}^s \,:\, i \in \mathcal{I}_j^s\right\}$ be the restriction of $\mathcal{U}^s$ to $\mathcal{I}_j^s$. We use $\mathcal{U}_0^s$ to estimate $Q_{f|s}$ and $\mathcal{U}_1^s$ to estimate $F_{f|s}$. For each $f : \mathbb{R}^d{\times}\mathcal{S} \to \mathbb{R}$ and each $s \in \mathcal{S}$, we estimate $\nu_{f|s}$ by

$$\hat{\nu}_{f|s}^0 = \frac{1}{|\mathcal{I}_0^s|} \sum_{i \in \mathcal{I}_0^s} \delta\big(f(X_i^s, s) + \varepsilon_{is} - \cdot\big) \quad \text{and} \quad \hat{\nu}_{f|s}^1 = \frac{1}{|\mathcal{I}_1^s|} \sum_{i \in \mathcal{I}_1^s} \delta\big(f(X_i^s, s) + \varepsilon_{is} - \cdot\big) \;, \tag{4}$$

where δ is the Dirac measure and all $\varepsilon_{is} \overset{\text{i.i.d.}}{\sim} U([-\sigma, \sigma])$, for some positive σ set by the user. Using the estimators in Eq. (4), we define for all $f : \mathbb{R}^d{\times}\mathcal{S} \to \mathbb{R}$ estimators of $Q_{f|s}$ and of $F_{f|s}$ as

$$\hat{Q}_{f|s} \equiv Q_{\hat{\nu}_{f|s}^0} \quad \text{and} \quad \hat{F}_{f|s} \equiv F_{\hat{\nu}_{f|s}^1} \;. \tag{5}$$

That is, $\hat{F}_{f|s}$ and $\hat{Q}_{f|s}$ are the empirical CDF and empirical quantiles of $(f(X, S){+}\varepsilon)|S{=}s$ based on $\{f(X_i^s, s){+}\varepsilon_{is}\}_{i \in \mathcal{I}_1^s}$ and $\{f(X_i^s, s){+}\varepsilon_{is}\}_{i \in \mathcal{I}_0^s}$ respectively. The noise ε_{is} serves as a smoothing random variable, since for all $s \in \mathcal{S}$ and $i \in [N_s]$ the random variables $f(X_i^s, s){+}\varepsilon_{is}$ are i.i.d. continuous for any $\mathbb{P}$ and f. In contrast, $f(X_i^s, s)$ might have atoms resulting in a non-zero probability to observe ties in $\{f(X_i^s, s)\}_{i \in \mathcal{I}_j^s}$. This step is also known as *jittering*, often used for data visualization [11] for tie-breaking. It plays a crucial role in the distribution-free fairness guarantees that we derive in Proposition 4.1; see the discussion thereafter.

Finally, let $\mathcal{A}{=}\{S_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_S$ and for each $s \in \mathcal{S}$ let $\hat{p}_s$ be the empirical frequency of $S{=}s$ evaluated on $\mathcal{A}$. Given a base estimator $\hat{f}$ of f^* constructed from n labeled samples $\mathcal{L}{=}\{(X_i, S_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, we define the final estimator $\hat{g}$ of g^* for all $(x, s) \in \mathbb{R}^d{\times}\mathcal{S}$ mimicking Eq. (3) as

$$\hat{g}(x, s) = \left(\sum_{s' \in \mathcal{S}} \hat{p}_{s'} \hat{Q}_{\hat{f}|s'}\right) \circ \hat{F}_{\hat{f}|s}\left(\hat{f}(x, s) + \varepsilon\right) \;, \tag{6}$$

[1] Since we are ready to sacrifice a factor 2 in our bounds, this assumption is without loss of generality.

Algorithm 1: Procedure to evaluate estimator in Eq. (6)

Input: new point: (x, s); base estimator $\hat{f}$; unlabeled data $\mathcal{U}^1, \ldots, \mathcal{U}^{|\mathcal{S}|}$; jitter parameter σ; empirical frequencies $\hat{p}_1, \ldots, \hat{p}_{|S|}$

Output : fair prediction $\hat{g}(x, s)$ for the point (x, s)

for $s' \in \mathcal{S}$ **do** // data structure for Eq. (4)

  $\mathcal{U}_0^{s'}, \mathcal{U}_1^{s'} \leftarrow \texttt{split\_in\_two}(\mathcal{U}^{s'})$ // split unlabeled data into two equal parts

  $\texttt{ar}_0^{s'} \leftarrow \left\{\hat{f}(X, s') + U([-\sigma, \sigma])\right\}_{X \in \mathcal{U}_0^{s'}}$, $\texttt{ar}_1^{s'} \leftarrow \left\{\hat{f}(X, s') + U([-\sigma, \sigma])\right\}_{X \in \mathcal{U}_1^{s'}}$

  $\texttt{ar}_0^{s'} \leftarrow \texttt{sort}\left(\texttt{ar}_0^{s'}\right)$, $\texttt{ar}_1^{s'} \leftarrow \texttt{sort}\left(\texttt{ar}_1^{s'}\right)$ // for fast evaluation of Eq. (5)

end

$k_s \leftarrow \texttt{position}\left(\hat{f}(x, s) + U([-\sigma, \sigma]),\ \texttt{ar}_1^s\right)$ // evaluate $\hat{F}_{\hat{f}|s}\left(\hat{f}(x, s) + \varepsilon\right)$ in Eq. (6)

$\hat{g}(x, s) \leftarrow \sum_{s' \in \mathcal{S}} \hat{p}_{s'} \times \texttt{ar}_0^{s'}\left[\lceil N_{s'} k_s / N_s \rceil\right]$ // evaluation of Eq. (6)

where $\varepsilon \sim U([-\sigma, \sigma])$ is assumed to be independent from every other random variables.

Remark 3.1. *In practice one should use a very small value for σ* (e.g., σ=10^{-5})*, which does not alter the statistical quality of the base estimator $\hat{f}$ as indicated in Theorem 4.4.*

A pseudo-code implementation of $\hat{g}$ in Eq. (6) is reported in Algorithm 1. It requires two primitives: $\texttt{sort}(\texttt{ar})$ sorts the array $\texttt{ar}$ in an increasing order; $\texttt{position}(a, \texttt{ar})$ which outputs the index k such that the insertion of a into k'th position in $\texttt{ar}$ preserves ordering (*i.e.,* $\texttt{ar}[k-1] \leq a < \texttt{ar}[k]$). Algorithm 1 consists of two for parts: in the for-loop we perform a preprocessing which takes $\sum_{s \in \mathcal{S}} O(N_s \log N_s)$ time[2] since it involves sorting; then, the evaluation of $\hat{g}$ on a new point (x, s) is performed in $(\max_{s \in \mathcal{S}} \log N_s)$ time since it involves an element search in a sorted array. Note that the for-loop of Algorithm 1 needs to be performed only once as this step is shared for any new (x, s).

4 Statistical analysis

In this section we provide a statistical analysis of the proposed algorithm. We first present in Proposition 4.1 distribution-free finite sample fairness guarantees for post-processing of *any* base learner with unlabeled data and then we show in Theorem 4.4 that if the base estimator $\hat{f}$ is a good proxy for f^*, then under mild assumptions on the distribution $\mathbb{P}$, the processed estimator $\hat{g}$ in Eq. (6) is a good estimator of g^* in Eq. (3).

Distribution free post-processing fairness guarantees. We derive two distribution-free results in Proposition 4.1, the first in Eq. (7) shows that the fairness definition is satisfied as long as we take the expectation over the data inside the supremum in Definition 2.2, while the second one in Eq. (8) bounds the expected violation of Definition 2.2.

Proposition 4.1 (Fairness guarantees)**.** *For any joint distribution $\mathbb{P}$ of (X, S, Y), any base estimator $\hat{f}$ constructed on labeled data, and for all $s, s' \in \mathcal{S}$, the estimator $\hat{g}$ defined in Eq. (6) satisfies*

$$\sup_{t \in \mathbb{R}} \left|\mathbf{P}(\hat{g}(X, S) \leq t | S{=}s) - \mathbf{P}(\hat{g}(X, S) \leq t | S{=}s')\right| \leq 2\left(N_s \wedge N_{s'} + 2\right)^{-1} \mathbf{1}_{\{N_s \neq N_{s'}\}} \tag{7}$$

$$\mathbf{E} \sup_{t \in \mathbb{R}} \left|\mathbf{P}(\hat{g}(X, S) \leq t | S{=}s, \mathcal{D}) - \mathbf{P}(\hat{g}(X, S) \leq t | S{=}s', \mathcal{D})\right| \leq 6\left(N_s \wedge N_{s'} + 1\right)^{-1/2} . \tag{8}$$

where $\mathcal{D} = \mathcal{L} \cup \mathcal{A} \cup_{s \in \mathcal{S}} \mathcal{U}^s$ is the union of all available datasets.

Let us point out that this result does not require any assumption on the distribution $\mathbb{P}$ as well as on the base estimator $\hat{f}$. This is achieved thanks to the jittering step in the definition of $\hat{g}$ in Eq. (6), which artificially introduces continuity. Continuity allows us to use results from the theory of rank statistics of exchangeable random variables to derive Eq. (7) as well as the classical inverse transform (see *e.g.,* [40, Sections 13 and 21]) combined with the Dvoretzky-Kiefer-Wolfowitz inequality [33]

[2] It is assumed in this discussion that the time complexity to evaluate $\hat{f}$ is $O(1)$.

to derive Eq. (8). Since basic results on rank statistics and inverse transform are distribution-free as long as the underlying random variable is continuous, the guarantees in Eqs. (7)–(8) are also distribution-free and can be applied on top of *any* base estimator $\hat{f}$.
The bound in Eq. (7) might be surprising to the reader. Yet, let us emphasize that this bound holds because the expectation *w.r.t.* the data distribution is taken inside the supremum (since $\mathbf{P}$ stands for the joint distribution of *all* random variables involved in $\hat{g}(X, S)$). Similar proof techniques are also used in randomization inference via permutations [19, 22], conformal prediction [30, 42], knockoff estimation [4] to name a few. However, unlike the aforementioned contributions, the problem of fairness requires a non-trivial adaptation of these techniques. In contrast, Eq. (8) might be more appealing to the machine learning community as it controls the expected (over data) violation of the fairness constraint with standard parametric rate.

Estimation guarantee with accurate base estimator. In order to prove non-asymptotic risk bounds we require the following assumption on the distribution $\mathbb{P}$ of $(X, S, Y) \in \mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$.

Assumption 4.2. *For each $s \in \mathcal{S}$ the univariate measure $\nu_{f^*|s}$ admits a density q_s, which is lower bounded by $\underline{\lambda}_s > 0$ and upper-bounded by $\overline{\lambda}_s \geq \underline{\lambda}_s$.*

Although the lower bound on the density assumption is rather strong and might potentially be violated in practice, it is still reasonable in certain situations. We believe that it can be replaced by the assumption that $f^*(X, S)$ conditionally on $S{=}s$ for all $s{\in}\mathcal{S}$ admits $2{+}\epsilon$ moments. We do not explore this relaxation in our work as it significantly complicates the proof of Theorem 4.4. At the same time, our empirical study suggests that the lower bound on the density is not intrinsic to the problem, since the estimator exhibits a good performance across various scenarios. In contrast, the milder assumption that the density is upper bounded is crucial for our proof and seems to be necessary.

Apart from the assumption on the density of $\nu_{f^*|s}$, the actual rate of estimation depends on the quality of the base estimator $\hat{f}$. We require the following assumption, which states that $\hat{f}$ approximates f^* point-wise with rate $b_n^{-1/2}$ and a standard sub-Gaussian concentration for $\hat{f}$ can be derived.

Assumption 4.3. *There exist positive constants c and C independent from n, N, $N_1, \dots, N_{|\mathcal{S}|}$, and a positive sequence $b_n : \mathbb{N} \to \mathbb{R}_+$ such that for all $\delta > 0$ it holds that*

$$\mathbf{P}\left(|f^*(x,s) - \hat{f}(x,s)| \geq \delta\right) \leq c \exp\left(-C b_n \delta^2\right) \text{ for almost all } (x,s) \text{ w.r.t. } \mathbb{P}_{X,S} \ .$$

We refer to [3, 15, 29, 30, 39] for various examples of estimators and additional assumptions such that the bound in Assumption 4.3 is satisfied. It includes local polynomial estimators, k-nearest neighbours, and linear regression, to name just a few.

Under these assumptions we can prove the following finite-sample estimation bound.

Theorem 4.4 (Estimation guarantee)**.** *Let Assumptions 4.2 and 4.3 be satisfied, and set $\sigma \lesssim \min_{s\in\mathcal{S}} N_s^{-1/2} \wedge b_n^{-1/2}$, then the estimator $\hat{g}$ defined in Eq. (6) satisfies*

$$\mathbf{E}\,|g^*(X,S) - \hat{g}(X,S)| \lesssim b_n^{-1/2} \bigvee \left(\sum_{s\in\mathcal{S}} p_s N_s^{-1/2}\right) \bigvee \sqrt{\frac{|\mathcal{S}|}{N}} \ ,$$

where the leading constant depends only on $\underline{\lambda}_s, \overline{\lambda}_s, C, c$ from Assumptions 4.2 and 4.3.

The proof of this result combines expected deviation of empirical measure from the real measure in terms of Wasserstein distance on real line [7] with the already mentioned rank statistics and classical peeling argument of [3].
The first term of the derived bound corresponds to the estimation error of f^* by $\hat{f}$, the second term is the price to pay for not knowing conditional distributions $X|S = s$ while the last term correspond to the price of unknown marginal probabilities of each protected attribute. Notice that if $N_s = p_s N$, which corresponds to the standard i.i.d. sampling from $\mathbb{P}_{X,S}$ of unlabeled data, the second and the third term are of the same order. Moreover, if N is sufficiently large, which in most scenarios[3] is

[3] One can achieve it by splitting the labeled dataset $\mathcal{L}$ artificially augmenting the unlabeled one, which ensures that $N > n$. In this case if $b_n^{-1/2} = O(n^{-1/2})$, then the first term is always dominant in the derived bound.

w.l.o.g., then the rate is dominated by $b_n^{-1/2}$. Notice that one can find a collection of joint distributions $\mathbb{P}$, such that f^* satisfies demographic parity. Hence, if $b_n^{-1/2}$ is the minimax optimal estimation rate of f^*, then it is also optimal for $g^* \equiv f^*$.

5 Empirical study

In this section, we present numerical experiments[4] with the proposed fair regression estimator defined in Section 3. In all experiments, we collect statistics on the test set $\mathcal{T} = \{(X_i, S_i, Y_i)\}_{i=1}^{n_{\text{test}}}$. The empirical mean squared error (MSE) is defined as

$$\text{MSE}\,(g) = \frac{1}{n_{\text{test}}} \sum_{(X,S,Y)\in\mathcal{T}} (Y - g(X,S))^2 \ .$$

We also measure the violation of fairness constraint imposed by Definition 2.2 via the empirical Kolmogorov-Smirnov (KS) distance,

$$\text{KS}\,(g) = \max_{s,s'\in\mathcal{S}} \sup_{t\in\mathbb{R}} \left| \frac{1}{|\mathcal{T}^s|} \sum_{(X,S,Y)\in\mathcal{T}^s} \mathbf{1}_{\{g(X,S)\le t\}} - \frac{1}{|\mathcal{T}^{s'}|} \sum_{(X,S,Y)\in\mathcal{T}^{s'}} \mathbf{1}_{\{g(X,S)\le t\}} \right| \ ,$$

where for all $s\in\mathcal{S}$ we define the set $\mathcal{T}^s = \{(X,S,Y) \in \mathcal{T} \,:\, S=s\}$. For all datasets we split the data in two parts (70% train and 30% test), this procedure is repeated 30 times, and we report the average performance on the test set alongside its standard deviation. We employ the 2-steps 10-fold CV procedure considered by [17] to select the best hyperparameters with the training set. In the first step, we shortlist all the hyperparameters with MSE close to the best one (in our case, the hyperparameters which lead to 10% larger MSE *w.r.t.* the best MSE). Then, from this list, we select the hyperparameters with the lowest KS.

Methods. We compare our method (see Section 3) to different fair regression approaches for both linear and non-linear regression. In the case of linear models we consider the following methods: Linear RLS plus [6] (RLS+Berk), Linear RLS plus [34] (RLS+Oneto), and Linear RLS plus Our Method (RLS+Ours), where RLS is the abbreviation of Regularized Least Squares.
In the case of non-linear models we compare to the following methods. i) For Kernel RLS (KRLS): KRLS plus [34] (KRLS+Oneto), KRLS plus [35] (KRLS+Perez), KRLS plus Our Method (KRLS+Ours); ii) For Random Forests (RF): RF plus [36] (RF+Raff), RF plus [1][5] (RF+Agar), and RF plus Our Method (RF+Ours).
The hyperparameters of the methods are set as follows. For RLS we set the regularization hyperparameters $\lambda \in 10^{\{-4.5,-3.5,\cdots,3\}}$ and for KRLS we set $\lambda \in 10^{\{-4.5,-3.5,\cdots,3\}}$ and $\gamma \in 10^{\{-4.5,-3.5,\cdots,3\}}$. Finally, for RF we set to 1000 the number of trees and for the number of features to select during the tree creation we search in $\{d^{1/4}, d^{1/2}, d^{3/4}\}$.

Datasets. In order to analyze the performance of our methods and test it against the state-of-the-art alternatives, we consider five benchmark datasets, CRIME, LAW, NLSY, STUD, and UNIV, which are briefly described below:
Communities&Crime (CRIME) contains socio-economic, law enforcement, and crime data about communities in the US [37] with 1994 examples. The task is to predict the number of violent crimes per 10^5 population (normalized to $[0,1]$) with race as the protected attribute. Following [9], we made a binary sensitive attribute s as to the percentage of black population, which yielded 970 instances of $s{=}1$ with a mean crime rate 0.35 and 1024 instances of $s{=}{-}1$ with a mean crime rate 0.13.
Law School (LAW) refers to the Law School Admissions Councils National Longitudinal Bar Passage Study [44] and has 20649 examples. The task is to predict a students GPA (normalized to $[0,1]$) with race as the protected attribute (white versus non-white).
National Longitudinal Survey of Youth (NLSY) involves survey results by the U.S. Bureau of Labor Statistics that is intended to gather information on the labor market activities and other life events of several groups [8]. Analogously to [27] we model a virtual company's hiring decision assuming that the company does not have access to the applicants' academic scores. We set as target the person's GPA (normalized to $[0,1]$), with race as sensitive attribute
Student Performance (STUD), approaches 649 students achievement (final grade) in secondary

[4] The source of our method can be found at `https://www.link-anonymous.link`.
[5] We thank the authors for sharing a prototype of their code.

Method	CRIME		LAW		NLSY		STUD		UNIV	
	MSE	KS	MSE	KS	MSE	KS	MSE	KS	MSE	KS
RLS	.033±.003	.55±.06	.107±.010	.15±.02	.153±.016	.73±.07	4.77±.49	.50±.05	2.24±.22	.14±.01
RLS+Berk	.037±.004	.16±.02	.121±.013	.10±.01	.189±.019	.49±.05	5.28±.57	.32±.03	2.43±.23	.05±.01
RLS+Oneto	.037±.004	.14±.01	.112±.012	.07±.01	.156±.016	.50±.05	5.02±.54	.23±.02	2.44±.26	.05±.01
RLS+Ours	.041±.004	.12±.01	.141±.014	.02±.01	.203±.019	.09±.01	5.62±.52	.04±.01	2.98±.32	.02±.01
KRLS	.024±.003	.52±.05	.040±.004	.09±.01	.061±.006	.58±.06	3.85±.36	.47±.05	1.43±.15	.10±.01
KRLS+Oneto	.028±.003	.19±.02	.046±.004	.05±.01	.066±.007	.06±.01	4.07±.39	.18±.02	1.46±.13	.04±.01
KRLS+Perez	.033±.003	.25±.02	.048±.005	.04±.01	.065±.007	.08±.01	3.97±.38	.14±.02	1.50±.15	.06±.01
KRLS+Ours	.034±.004	.09±.01	.056±.005	.01±.01	.081±.008	.03±.01	4.46±.43	.03±.01	1.71±.16	.02±.01
RF	.020±.002	.45±.04	.046±.005	.11±.01	.055±.006	.55±.06	3.59±.39	.45±.05	1.31±.13	.10±.01
RF+Raff	.030±.003	.21±.02	.058±.006	.06±.01	.066±.006	.08±.01	4.28±.40	.09±.01	1.38±.12	.02±.01
RF+Agar	.029±.003	.13±.01	.050±.005	.04±.01	.065±.006	.07±.01	3.87±.41	.07±.01	1.40±.13	.02±.01
RF+Ours	.033±.003	.08±.01	.064±.006	.02±.01	.070±.007	.03±.01	4.18±.38	.02±.01	1.49±.14	.01±.01

Table 1: Results for all the datasets and all the methods concerning MSE and KS.

education of two Portuguese schools using 33 attributes [14], with gender as the protected attribute. *University Anonymous (UNIV)* is a proprietary and highly sensitive dataset containing all the data about the past and present students enrolled at the University of *Anonymous*. In this study we take into consideration students who enrolled, in the academic year 2017-2018. The dataset contains 5000 instances, each one described by 35 attributes (both numeric and categorical) about ethnicity, gender, financial status, and previous school experience. The scope is to predict the average grades at the end of the first semester, with gender as the protected attribute.

Comparison w.r.t. state-of-the-art. In Table 1, we present the performance of different methods on various datasets described above. One can notice that LAW and UNIV datasets have a least amount of disciminatory bias (quantified by KS), since the fairness *unaware* methods perform reasonably well in terms of KS. Furthermore, on these two datasets, the difference in performance between all fairness aware methods is less noticeable. In contrast, on CRIME, NLSY, and STUD, fairness unaware methods perform poorly in terms of KS. More importantly, our findings indicate that the proposed method is competitive with state-of-the-art methods and is the most effective in imposing the fairness constraint. In particular, in all except two considered scenarios (CRIME+RLS, CRIME+RF) our method improves fairness by 50% (and up to 80% in some cases) over the closest fairness aware method. In contrast, the accuracy of our method decreases by 1% up to 30% when compared to the most accurate fairness aware method. However, let us emphasize that the relative decrease in accuracy is much smaller than the relative improvement in fairness across the considered scenarios. For example, on NLSY+RLS the most accurate fairness aware method is RLS+Oneto with mean MSE$=.156$ and mean KS$=.50$, while RLS+Ours yields mean MSE$=.203$ and mean KS$=.09$. That is, compared to RLS+Oneto our method drops about 30% in accuracy, while gains about 82% in fairness. With RF, which is a more powerful estimator, the average drop in accuracy across all datasets compared to RF+Agar is about 12% while the average improvement in fairness is about 53%.

6 Conclusion and perspectives

In this work we investigated the problem of fair regression with Demographic Parity constraint assuming that the sensitive attribute is available for prediction. We derived a closed form solution for the optimal fair predictor which offers a simple and intuitive interpretation. Relying on this expression, we devised a post-processing procedure, which transforms any base estimator of the regression function into a nearly fair one, independently of the underlying distribution. Moreover, if the base estimator is accurate, our post-processing method yields an accurate estimator of the optimal fair predictor as well. Finally, we conducted an empirical study indicating the effectiveness of our method in imposing fairness in practice. In the future it would be valuable to extend our methodology to the case when we are not allowed to use the sensitive feature as well as to other notions of fairness.

7 Acknowledgement

This work was supported by the Amazon AWS Machine Learning Research Award, SAP SE, and CISCO.

Broader impact

This work investigates the problem of fair regression with multiple sensitive groups using tools from statistical learning theory and optimal transport theory. Our results lead to an efficient learning algorithm that we show empirically and theoretically to be very effective to impose fairness according to the notion of Demographic Parity. Our approach is directly designed to mitigate potential bias present in the data. Hence, even though the work is primarily theoretical, we anticipate that our results could be used in the future by practitioners in order to specialize our methodology to real-life scenarios involving individuals, and to potentially help making decision which help people with disadvantages or minority groups.

We believe that the most important positive impact of our work is the intuitive interpretation of the optimal fair prediction, which should help to reason as to why a given prediction was made for a given individual. At the same time, this interpretation allows to understand the weaknesses of the notion of Demographic Parity: if f^* does not adequately reflect the group-wise ordering of individuals, the optimal fair prediction g^* might not lead to a fair prediction from individuals' perspective. In other words, returning to the salary example considered above, the notion of Demographic Parity reflects the principle: more qualified individuals get higher salary within their respective groups.

References

[1] A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 2019.

[2] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] J. Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[4] Rina Foygel Barber, Emmanuel J Candès, et al. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018.

[6] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[7] S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and kantorovich transport distances. *Memoirs of the American Mathematical Society*, 2016.

[8] Bureau of Labor Statistics. National longitudinal surveys of youth data set. www.bls.gov/nls/, 2019.

[9] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *IEEE International Conference on Data Mining*, 2013.

[10] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*, 2017.

[11] J. M. Chambers. *Graphical methods for data analysis*. CRC Press, 2018.

[12] S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *AAAI*, 2020.

[13] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Neural Information Processing Systems*, 2017.

[14] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. In *FUture BUsiness TEChnology Conference*, 2008.

[15] L. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24(2):142–151, 1978.

[16] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018.

[17] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018.

[18] C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

[19] Ronald Aylmer Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936.

[20] P. Gordaliza, E. Del Barrio, G. Fabrice, and J. M. Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2019.

[21] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

[22] Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192, 1952.

[23] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fair learning in markovian environments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[24] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

[25] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016.

[26] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Neural Information Processing Systems*, 2017.

[27] J. Komiyama and H. Shimao. Comparing fairness criteria based on social outcome. *arXiv preprint arXiv:1806.05112*, 2018.

[28] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Neural Information Processing Systems*, 2017.

[29] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

[30] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

[31] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[32] K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

[33] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[34] L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019.

[35] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[36] E. Raff, J. Sylvester, and S. Mills. Fair forests: Regularized tree induction to minimize model bias. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[37] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[38] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[39] S. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[40] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[41] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[42] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[43] H. Wang, B. Ustun, and F. Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, 2019.

[44] L. F. Wightman and H. Ramsey. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[45] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Neural Information Processing Systems*, 2017.

[46] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

[47] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

[48] Gianluca Zeni, Matteo Fontana, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *arXiv preprint arXiv:2005.07972*, 2020.

[49] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Supplementary material

The supplementary material is organized as follows. In Appendix A we provide the proof of Theorem 2.3, in Appendix B we provide the proof of Proposition 4.1, and in Appendix C we prove Theorem 4.4. For reader's convenience all the results are repeated in this supplementary material and a short overview of classical results is provided.

A Characterization of the optimal

Before providing the proof of Theorem 2.3, let us give a brief overview of classical results in the Optimal transport theory with one dimensional measures; all the results can be found in [38, 41]

Definition 2.1 (Wasserstein-2 distance)**.** *Let μ and ν be two univariate probability measures. The squared Wasserstein-2 distance between μ and ν is defined as*

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma_{\mu,\nu}} \int |x-y|^2 \, d\gamma(x,y) \ ,$$

where $\Gamma_{\mu,\nu}$ is the set of distributions (couplings) on $\mathbb{R}\times\mathbb{R}$ such that for all $\gamma \in \Gamma_{\mu,\nu}$ and all measurable sets $A, B \subset \mathbb{R}$ it holds that $\gamma(A\times\mathbb{R}) = \mu(A)$ and $\gamma(\mathbb{R}\times B) = \nu(B)$.

The coupling γ which achieves the infimum in the definition of the Wasserstein-2 distance is called the optimal coupling.

Also let us mention that the Wasserstein-2 distance between two univariate probability measures ν, μ, defined in Definition 2.1, can be expressed as

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{\gamma} \mathbb{E}_{(Z_\mu,Z_\nu)\sim\gamma}(Z_\nu - Z_\mu)^2 \ ,$$

where $Z_\nu \sim \nu$ and $Z_\mu \sim \mu$ and the infimum is taken over all joint distributions γ of (Z_ν, Z_μ) which preserve marginals.

The next result establishes that as long as one of the measures in the definition of the Wasserstein-2 distance admits a density, then the optimal coupling in the infimum in Definition 2.1 is deterministic (see *e.g.,* [41, Theorem 2.18] or [38, Theorems 2.5 and 2.9]).

Theorem A.1. *Let ν, μ be two univariate measures such that ν has a density and let $X \sim \nu$. Then there exists a mapping $T : \mathbb{R} \to \mathbb{R}$ such*

$$\mathcal{W}_2^2(\mu,\nu) = \mathbb{E}(X - T(X))^2 \ ,$$

that is $(X, T(X)) \sim \bar\gamma \in \Gamma_{\mu,\nu}$ where $\bar\gamma$ is an optimal coupling. Moreover, the transport map is given by $T = Q_\mu \circ F_\nu$.

By the abuse of notation, for an increasing real-valued univariate function F we will use $F^{\leftarrow}$ to denote its generalized inverse. For instance, if $F : \mathbb{R} \to [0,1]$ is a CDF, then $F^{\leftarrow}$ is the quantile function that was defined in the introduction.

The next result is standard and can be found for instance in [2, Section 6.1] or [38, Section 5.5.5]. It states that for one dimensional Wasserstein barycenter problem, the optimal measure admits a closed form solution.

Lemma A.2. *Let $\nu_1, \ldots, \nu_{|\mathcal{S}|}$ be $|\mathcal{S}|$ univariate probability measures admitting densities, for all $p_1, \ldots, p_{|\mathcal{S}|} \geq 0$ such that $p_1 + \ldots + p_{|\mathcal{S}|} = 1$ define*

$$\nu^* \in \arg\min_{\nu} \sum_{s=1}^{|\mathcal{S}|} p_s \mathcal{W}_2^2(\nu_s, \nu) \ .$$

Then, the cumulative distribution of ν^* is given by

$$F_{\nu^*}(\cdot) = \left(\sum_{s=1}^{|\mathcal{S}|} p_s Q_{\nu_s} \right)^{\leftarrow} (\cdot) \ .$$

Theorem A.1 and Lemma A.2 are the two main ingredients that are used in the proof of Theorem 2.3.

Theorem 2.3 (Characterization of fair optimal prediction)**.** *Assume, for each $s \in \mathcal{S}$, that the univariate measure $\nu_{f^*|s}$ has a density and let $p_s = \mathbb{P}(S = s)$. Then,*

$$\min_{g \text{ is fair}} \mathbb{E}(f^*(X,S) - g(X,S))^2 = \min_{\nu} \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \ .$$

Moreover, if g^* and ν^* solve the l.h.s. and the r.h.s. problems respectively, then $\nu^* = \nu_{g^*}$ and

$$g^*(x,s) = \left(\sum_{s'\in\mathcal{S}} p_{s'} Q_{f^*|s'} \right) \circ F_{f^*|s} \left(f^*(x,s) \right) \ . \tag{3}$$

Proof of Theorem 2.3. We want to show that

$$\min_{g \text{ is fair}} \mathbb{E}(f^*(X,S) - g(X,S))^2 = \min_{\nu} \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \ .$$

Let $\bar{g} : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ be a minimizer of the *l.h.s.* of the above equation and define by $\nu_{\bar{g}}$ the distribution of $\bar{g}$. Since $\nu_{f^*|s}$ admits density, using Theorem A.1 for each $s \in \mathcal{S}$ there exists $T_s = Q_{\nu_{\bar{g}}} \circ F_{f^*|s}$ such that

$$\begin{aligned}
\sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu_{\bar{g}}) &= \sum_{s\in\mathcal{S}} p_s \int_{\mathbb{R}} (z - T_s(z))^2 \, d\nu_{f^*|s}(z) \\
&= \sum_{s\in\mathcal{S}} p_s \int_{\mathbb{R}^d} (f^*(x,s) - T_s \circ f^*(x,s))^2 \, d\mathbb{P}_{X|S=s}(x) \\
&= \sum_{s\in\mathcal{S}} p_s \mathbb{E}\left[(f^*(X,s) - (T_s \circ f^*)(X,s))^2 \, | S = s \right] \\
&= \mathbb{E}(f^*(X,S) - \tilde{g}(X,S))^2 \ ,
\end{aligned}$$

where we defined $\tilde{g}$ for all $(x,s) \in \mathbb{R}^d \times \mathcal{S}$ as

$$\tilde{g}(x,s) = (T_s \circ f^*)(x,s) = \left(Q_{\nu_{\bar{g}}} \circ F_{f^*|s} \circ f^* \right)(x,s) \ .$$

The cumulative distribution of $\tilde{g}$ can be expressed as

$$\begin{aligned}
\mathbb{P}(\tilde{g}(X,S) \le t) &= \sum_{s\in\mathcal{S}} p_s \mathbb{P}_{X|S=s} \left(Q_{\nu_{\bar{g}}} \circ F_{f^*|s} \circ f^*(X,s) \le t \right) \\
&= \sum_{s\in\mathcal{S}} p_s \mathbb{P}_{X|S=s} \left(f^*(X,s) \le Q_{f^*|s} \circ F_{\nu_{\bar{g}}}(t) \right) = F_{\nu_{\bar{g}}}(t) \ ,
\end{aligned}$$

where the last equality is due to the fact that $\nu_{f^*|s}$ admits a density for all $s \in \mathcal{S}$. The above implies that $\tilde{g}$ is fair, thus on the one hand by optimality of $\bar{g}$ we have

$$\mathbb{E}(f^*(X,S) - \tilde{g}(X,S))^2 \ge \mathbb{E}(f^*(X,S) - \bar{g}(X,S))^2 \ ,$$

on the other hand we have for each $s \in \mathcal{S}$

$$\mathcal{W}_2^2(\nu_{f^*|s}, \nu_{\bar{g}}) \le \mathbb{E}\left[(f^*(X,S) - \bar{g}(X,S))^2 \, | S = s \right] \ .$$

Thus we showed that

$$\sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu_{\bar{g}}) = \min_{g \text{ is fair}} \mathbb{E}(f^*(X,S) - g(X,S))^2 \ . \tag{9}$$

This implies that

$$\min_{\nu} \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \le \min_{g \text{ is fair}} \mathbb{E}(f^*(X,S) - g(X,S))^2 \ . \tag{10}$$

Now we want to show that the opposite inequality also holds. To this end define ν^* as

$$\nu^* \in \arg\min_{\nu} \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \ .$$

Set T_s^* as optimal transport maps from $\nu_{f^*|s}$ to ν^* of the form $T_s^* = Q_{\nu^*} \circ F_{f^*|s}$ (provided by Theorem A.1 and our assumption on the density of $\nu_{f^*|s}$) and define g^* for all $(x, s) \in \mathbb{R}^d \times \mathcal{S}$ as

$$g^*(x, s) = \left(Q_{\nu^*} \circ F_{f^*|s} \circ f^*\right)(x, s) \ . \tag{11}$$

By the definition of g^* in Eq. (11) and Theorem A.1 we clearly have

$$\min_{\nu} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) = \mathbb{E}(f^*(X, S) - g^*(X, S))^2 \ . \tag{12}$$

Moreover since ν^* is independent from S, using similar argument as above one can show that g^* satisfies the Demographic Parity constraint in Definition 2.2 and thus, Eq. (12) yields

$$\min_{\nu} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \geq \min_{g \text{ is fair}} \mathbb{E}(f^*(X, S) - g(X, S))^2 \ . \tag{13}$$

Eqs. (10) and (13) yield the first assertion of the result. Notice that thanks to Eq. (12) we have also shown that

$$\mathbb{E}(f^*(X, S) - g^*(X, S))^2 = \mathbb{E}(f^*(X, S) - \bar{g}(X, S))^2 \ ,$$

and since g^* is fair we can put $\bar{g} = g^*$. Finally, using Lemma A.2 we derive an explicit form of ν^* and conclude using Eq. (11). □

B Proof of Proposition 4.1

Let us first recall the well-known Dvoretzky–Kiefer–Wolfowitz inequality [33, Corollary 1].

Theorem B.1 (Dvoretzky–Kiefer–Wolfowitz inequality)**.** *Let $Z_1, \ldots, Z_n$ be* i.i.d. *real valued random variables with cumulative distribution F. Let $\hat{F}$ be the empirical cumulative distribution of $Z_1, \ldots, Z_n$, then*

$$\mathbf{E}\|F - \hat{F}\|_\infty := \mathbf{E} \sup_{t \in \mathbb{R}} |F(t) - \hat{F}(t)| \leq \sqrt{\frac{\pi}{2n}} \ .$$

Proposition 4.1 (Fairness guarantees)**.** *For any joint distribution $\mathbb{P}$ of (X, S, Y), any base estimator $\hat{f}$ constructed on labeled data, and for all $s, s' \in \mathcal{S}$, the estimator $\hat{g}$ defined in Eq. (6) satisfies*

$$\sup_{t \in \mathbb{R}} |\mathbf{P}(\hat{g}(X, S) \leq t | S{=}s) - \mathbf{P}(\hat{g}(X, S) \leq t | S{=}s')| \leq 2 \left(N_s \wedge N_{s'} + 2\right)^{-1} \mathbf{1}_{\{N_s \neq N_{s'}\}} \tag{7}$$

$$\mathbf{E} \sup_{t \in \mathbb{R}} |\mathbf{P}(\hat{g}(X, S) \leq t | S{=}s, \mathcal{D}) - \mathbf{P}(\hat{g}(X, S) \leq t | S{=}s', \mathcal{D})| \leq 6 \left(N_s \wedge N_{s'} + 1\right)^{-1/2} \ . \tag{8}$$

where $\mathcal{D} = \mathcal{L} \cup \mathcal{A} \cup_{s \in \mathcal{S}} \mathcal{U}^s$ is the union of all available datasets.

Proof of Proposition 4.1. The proof of Eq. (7) is based on standard results in the theory of rank statistics (see e.g. [40, Sec. 13]). Meanwhile, the proof of Eq. (8) is built upon the well-known Dvoretzky–Kiefer–Wolfowitz inequality [33, Corollary 1].

Notice that if $X^s \sim \mathbb{P}_{X|S=s}$ and X^s is independent from labeled, unlabeled data, and the noise variables $\varepsilon_{is}, \varepsilon$, then it holds that

$$\mathbf{P}(\hat{g}(X, S) \leq t | S = s) = \mathbf{P}(\hat{g}(X^s, s) \leq t), \quad \forall t \in \mathbb{R} \ .$$

Proof of Eq. (7): We have for all $s, s' \in \mathcal{S}$ that

$$\begin{aligned} &\sup_{t \in \mathbb{R}} \left| \mathbf{P}(\hat{g}(X^s, s) \leq t) - \mathbf{P}(\hat{g}(X^{s'}, s') \leq t) \right| \\ &\quad \leq \sup_{t \in (0,1)} \left| \mathbf{P}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s, s) + \varepsilon\right) \leq t\right) - \mathbf{P}\left(\hat{F}_{\hat{f}|s'}\left(\hat{f}(X^{s'}, s') + \varepsilon\right) \leq t\right) \right| \ , \end{aligned}$$

where, thanks to the form of $\hat{g}$ in Eq. (6), the inequality follows from the fact that for all $s \in \mathcal{S}$

$$\left(\sum_{\tilde{s} \in \mathcal{S}} \hat{p}_{\tilde{s}} \hat{Q}_{\hat{f}|\tilde{s}}\right) \circ \hat{F}_{\hat{f}|s}\left(\hat{f}(x, s) + \varepsilon\right) \leq t \quad \Leftrightarrow \quad \hat{F}_{\hat{f}|s}\left(\hat{f}(x, s) + \varepsilon\right) \leq \left(\sum_{\tilde{s} \in \mathcal{S}} \hat{p}_{\tilde{s}} \hat{Q}_{\hat{f}|\tilde{s}}\right)^{\leftarrow}(t) \ .$$

Fix some $t \in (0,1)$ and let $k_s(t) \in \{1, \dots, |\mathcal{I}_1^s|\}$ be such that $\frac{k_s(t)-1}{|\mathcal{I}_1^s|} \leq t < \frac{k_s(t)}{|\mathcal{I}_1^s|}$, then by the definition of $\hat{F}_{\hat{f}|s}(\cdot)$ we have

$$\hat{F}_{\hat{f}|s}\left(\hat{f}(x,s)+\varepsilon\right) \leq t \quad \Leftrightarrow \quad \sum_{i \in \mathcal{I}_1^s} \mathbf{1}_{\left\{\hat{f}(X_i^s,s)+\varepsilon_{is} \leq \hat{f}(x,s)+\varepsilon\right\}} \leq k_s(t)-1 \ .$$

Notice that the random variables $\{\hat{f}(X^s,s)+\varepsilon\} \cup \{\hat{f}(X_i^s,s)+\varepsilon_{is}\}_{i \in \mathcal{I}_1^s}$ conditionally on labeled data $\mathcal{L}$ are i.i.d. and continuous. Thus, conditionally on $\mathcal{L}$ the random variable $\sum_{i \in \mathcal{I}_1^s} \mathbf{1}_{\left\{\hat{f}(X_i^s,s)+\varepsilon_{is} \leq \hat{f}(X^s,s)+\varepsilon\right\}}$ is distributed uniformly on $\{0, \dots, |\mathcal{I}_1^s|\}$ (see *e.g.,* [40, Lemma 13.1]), so that

$$\mathbf{P}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right) \leq t\right) = \frac{k_s(t)}{|\mathcal{I}_1^s|+1} \ .$$

Repeating the same argument for s' and recalling that $|\mathcal{I}_1^s| = N_s/2$ and $|\mathcal{I}_1^{s'}| = N_{s'}/2$, we get

$$\begin{aligned}\sup_{t \in \mathbb{R}} \left|\mathbf{P}(\hat{g}(X^s,s) \leq t) - \mathbf{P}(\hat{g}(X^{s'},s') \leq t)\right| &\leq \sup_{t \in (0,1)} \left|\frac{k_s(t)}{N_s/2+1} - \frac{k_{s'}(t)}{N_{s'}/2+1}\right| \\ &= 2\left(N_s \wedge N_{s'}+2\right)^{-1} \mathbf{1}_{\{N_s \neq N_{s'}\}} \ .\end{aligned}$$

Proof of Eq. (8): Similarly, as in the proof of Eq. (7) we can write

$$\begin{aligned}(*) &= \sup_{t \in \mathbb{R}} \left|\mathbf{P}(\hat{g}_s(X^s) \leq t|\mathcal{D}) - \mathbf{P}(\hat{g}_{s'}(X^{s'}) \leq t|\mathcal{D})\right| \\ &\leq \sup_{t \in (0,1)} \left|\mathbf{P}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right) \leq t|\mathcal{D}\right) - \mathbf{P}\left(\hat{F}_{\hat{f}|s'}\left(\hat{f}(X^{s'},s')+\varepsilon\right) \leq t|\mathcal{D}\right)\right| \ .\end{aligned}$$

Moreover, thanks to the triangle inequality we have

$$\begin{aligned}(*) &\leq \sup_{t \in (0,1)} \left|\mathbf{P}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right) \leq t|\mathcal{D}\right) - \mathbf{P}\left(F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right) \leq t|\mathcal{D}\right)\right| \\ &\quad + \sup_{t \in (0,1)} \left|\mathbf{P}\left(\hat{F}_{\hat{f}|s'}\left(\hat{f}(X^{s'},s')+\varepsilon\right) \leq t|\mathcal{D}\right) - \mathbf{P}\left(F_{\bar{\nu}_{\hat{f}|s'}}\left(\hat{f}(X^{s'},s')+\varepsilon\right) \leq t|\mathcal{D}\right)\right| \\ &= \sup_{t \in (0,1)} A_s(t) + \sup_{t \in (0,1)} A_{s'}(t) \ , \end{aligned} \tag{14}$$

where for all $t \in \mathbb{R}$ and all $s \in \mathcal{S}$ we defined

$$F_{\bar{\nu}_{\hat{f}|s}}(t) = \mathbf{P}\left(\hat{f}(X^s,s)+\varepsilon \leq t|\mathcal{D}\right) \ ,$$

and we used the fact that $\hat{f}(X^s,s)+\varepsilon$ is continuous conditionally on all the available data $\mathcal{D}$, then the random variable $F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right)$ is distributed uniformly on $(0,1)$ (see *e.g.,* [40, Lemma 21.1]), which means that for all $t \in (0,1)$ and all $s, s' \in \mathcal{S}$

$$t = \mathbf{P}\left(F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right) \leq t|\mathcal{D}\right) = \mathbf{P}\left(F_{\bar{\nu}_{\hat{f}|s'}}\left(\hat{f}(X^{s'},s')+\varepsilon\right) \leq t|\mathcal{D}\right) \ .$$

We bound the first term in Eq. (14) and the bound for the second terms follows the same arguments. Fix some $t \in (0,1)$, then we can write

$$\begin{aligned}A_s(t) &\leq \mathbf{P}\left(\left|F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right) - t\right| \leq \left|F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right) - \hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right| \middle| \mathcal{D}\right) \\ &\leq \mathbf{P}\left(\left|F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right) - t\right| \leq \left\|F_{\bar{\nu}_{\hat{f}|s}} - \hat{F}_{\hat{f}|s}\right\|_\infty \middle| \mathcal{D}\right) \leq 2\left\|F_{\bar{\nu}_{\hat{f}|s}} - \hat{F}_{\hat{f}|s}\right\|_\infty \ .\end{aligned}$$

Taking supremum on both sides and repeating the same argument for s', we get

$$(*) \leq 2\mathbf{E}\left\|F_{\bar{\nu}_{\hat{f}|s}} - \hat{F}_{\hat{f}|s}\right\|_\infty + 2\mathbf{E}\left\|F_{\bar{\nu}_{\hat{f}|s'}} - \hat{F}_{\hat{f}|s'}\right\|_\infty \ ,$$

we conclude applying Dvoretzky–Kiefer–Wolfowitz inequality, recalled in Theorem B.1, conditionally on $\mathcal{L}$. □

C Proof of Theorem 4.4

Let us first recall the assumptions that we require in order to prove Theorem 4.4.

Assumption 4.2. *For each $s \in \mathcal{S}$ the univariate measure $\nu_{f^*|s}$ admits a density q_s, which is lower bounded by $\underline{\lambda}_s > 0$ and upper-bounded by $\overline{\lambda}_s \geq \underline{\lambda}_s$.*

Assumption 4.3. *There exist positive constants c and C independent from n, N, $N_1, \dots, N_{|\mathcal{S}|}$, and a positive sequence $b_n : \mathbb{N} \to \mathbb{R}_+$ such that for all $\delta > 0$ it holds that*

$$\mathbf{P}\left(|f^*(x,s) - \hat{f}(x,s)| \geq \delta\right) \leq c \exp\left(-C b_n \delta^2\right) \text{ for almost all } (x,s) \text{ w.r.t. } \mathbb{P}_{X,S} \ .$$

The next simple result states that Assumption 4.3 yields a bound in L_1-norm between f^* and $\hat{f}$.

Lemma C.1. *Let Assumption 4.3 be satisfied, then for all $s \in \mathcal{S}$ it holds that*

$$\mathbf{E}\left[|f^*(X,S) - \hat{f}(X,S)||S = s\right] \leq \mathtt{A} b_n^{-1/2} \ ,$$

with $\mathtt{A} = \frac{c}{2}\sqrt{\frac{\pi}{C}}$.

Proof. Applying Fubini's theorem we can write

$$\begin{aligned}
\mathbf{E}\left[|f^*(X,S) - \hat{f}(X,S)||S = s\right] &= \int_{x \in \mathbb{R}^d} \mathbf{E}|f^*(x,s) - \hat{f}(x,s)| \mathbb{P}_{X|S=s}(dx) \\
&= \int_{x \in \mathbb{R}^d} \left(\int_0^{+\infty} \mathbf{P}(|f^*(x,s) - \hat{f}(x,s)| > t) dt \right) \mathbb{P}_{X|S=s}(dx) \\
&\overset{(a)}{\leq} \int_{x \in \mathbb{R}^d} \left(\int_0^{+\infty} c \exp\left(-C b_n t^2\right) dt \right) \mathbb{P}_{X|S=s}(dx) \\
&= c \int_0^{+\infty} \exp\left(-C b_n t^2\right) dt \ .
\end{aligned}$$

where (a) follows from Assumption 4.3. Making change of variables we get

$$c \int_0^{+\infty} \exp\left(-C b_n t^2\right) dt = c (C b_n)^{-1/2} \int_0^{+\infty} \exp\left(-t^2\right) dt = c (C b_n)^{-1/2} \frac{\sqrt{\pi}}{2} \ .$$

□

We also need to define Wasserstein 1 and ∞ distances.

Definition C.2. *Let μ and ν be two univariate probability measures, then Wasserstein 1 and ∞ distance between μ and ν are defined as*

$$\mathcal{W}_1(\mu,\nu) = \int_0^1 |Q_\mu(t) - Q_\nu(t)| \, dt \quad \text{and} \quad \mathcal{W}_\infty(\mu,\nu) = \sup_{t \in [0,1]} |Q_\mu(t) - Q_\nu(t)| \ ,$$

respectively.

Remark C.3. *The definitions of $\mathcal{W}_1$ and $\mathcal{W}_\infty$ can be stated in terms of couplings as it is done in Definition 2.1. However, for our purposes it is more convenient to use their equivalent formulation stated in Definition C.2. We refer to [7] and in particular to their Theorem 2.10 for further details.*

The final ingredient is [7, Theorem 5.11].

Theorem C.4. *Let $Z_1, \dots, Z_n$ be* i.i.d. *real valued random variables from some probability measure μ and let $\hat{\mu}$ be the empirical measure based on $Z_1, \dots, Z_n$. Assume that μ admits density which is lower-bounded by some constant $L > 0$, then*

$$\mathbf{E}[\mathcal{W}_\infty(\mu, \hat{\mu})] \leq L^{-1} \sqrt{\frac{2\pi}{n}} \ .$$

Theorem 4.4 (Estimation guarantee). *Let Assumptions 4.2 and 4.3 be satisfied, and set* $\sigma \lesssim \min_{s\in\mathcal{S}} N_s^{-1/2} \wedge b_n^{-1/2}$, *then the estimator* $\hat{g}$ *defined in Eq. (6) satisfies*

$$\mathbf{E}\left|g^*(X,S) - \hat{g}(X,S)\right| \lesssim b_n^{-1/2} \bigvee \left(\sum_{s\in\mathcal{S}} p_s N_s^{-1/2}\right) \bigvee \sqrt{\frac{|\mathcal{S}|}{N}} \ ,$$

where the leading constant depends only on $\underline{\lambda}_s, \overline{\lambda}_s, C, c$ *from Assumptions 4.2 and 4.3.*

Proof of Theorem 4.4. In the proof $a > 0$ is going to denote an absolute constant independent from the size of data, which can differ from line to line. First of all, define the random variable

$$\Delta(\hat{g}) = \mathbb{E}\left|\hat{g}(X,S) - g^*(X,S)\right| = \sum_{s\in\mathcal{S}} p_s \mathbb{E}\left[\left|\hat{g}(X,s) - g^*(X,s)\right| | S = s\right] \ ,$$

where $\mathbb{E}$ stands for the expectation *w.r.t.* the joint distribution of (X,S,Y). Recall that

$$g^*(x,s) = \sum_{s'\in\mathcal{S}} p_{s'} Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(x,s)\right)\right) \text{ and } \hat{g}(x,s) = \sum_{s'\in\mathcal{S}} \hat{p}_{s'} \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(x,s)+\varepsilon\right)\right) \ .$$

Considering $g^*(x,s)$ first, we can state that

$$\left|g^*(x,s) - \sum_{s'\in\mathcal{S}} \hat{p}_{s'} Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(x,s)\right)\right)\right| \leq \sum_{s'\in\mathcal{S}} |p_{s'} - \hat{p}_{s'}| \times \left|Q_{f^*|s'} \circ F_{f^*|s} \circ f^*(x,s)\right| \ .$$

It is clear that if we can find a bound on $|f^*(x,s')|$ which holds for almost all x *w.r.t.* $\mathbb{P}_{X|S=s'}$, it would imply an upper bound on $|Q_{f^*|s'}(t)|$ for all $t \in [0,1]$. Fix some $a > 0$, then on the one hand for all $s' \in \mathcal{S}$

$$\mathbb{P}(|f^*(X,S)| \leq a | S = s') \leq 1 \ ,$$

on the other hand under Assumption 4.2 we can write for all $s' \in \mathcal{S}$

$$\mathbb{P}(|f^*(X,S)| \leq a | S{=}s') = \int_{|f^*(x,s')|\leq a} \mathbb{P}_{X|S=s'}(dx) = \int_{|t|\leq a} q_{s'}(t)dt \geq \underline{\lambda}_{s'} \int_{|t|\leq a} dt = 2a\underline{\lambda}_{s'} \ ,$$

which implies that $a \leq 1/(2\underline{\lambda}_{s'})$ and therefore $|f^*(x,s')| \leq 1/(2\underline{\lambda}_{s'})$ for almost all $x \in \mathbb{R}^d$ *w.r.t.* $\mathbb{P}_{X|S=s'}$. Hence, we can write for all $(x,s) \in \mathbb{R}^d \times \mathcal{S}$

$$\left|g^*(x,s) - \sum_{s'\in\mathcal{S}} \hat{p}_{s'} Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(x,s)\right)\right)\right| \leq \frac{1}{2} \sum_{s'\in\mathcal{S}} \underline{\lambda}_{s'}^{-1} |p_{s'} - \hat{p}_{s'}| \ .$$

The above implies that

$$\begin{aligned}\Delta(\hat{g}) \leq & \sum_{s\in\mathcal{S}} p_s \sum_{s'\in\mathcal{S}} \hat{p}_{s'} \mathbb{E}\left[\left|Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(X,S)\right)\right) - \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X,s)+\varepsilon\right)\right)\right| | S = s\right] \\ & + \frac{1}{2} \sum_{s\in\mathcal{S}} \underline{\lambda}_s^{-1} |p_s - \hat{p}_s| \ .\end{aligned}$$

Taking the total expectation we arrive at

$$\begin{aligned}\mathbf{E}[\Delta(\hat{g})] \leq & \sum_{s,s'\in\mathcal{S}} p_s p_{s'} \mathbf{E}\left[\left|Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(X,S)\right)\right) - \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X,S)+\varepsilon\right)\right)\right| | S = s\right] \\ & + \frac{1}{2} \sum_{s\in\mathcal{S}} \underline{\lambda}_s^{-1} \mathbf{E}|p_s - \hat{p}_s| \ ,\end{aligned}$$

where we used the fact that $\hat{p}_s$ is an unbiased estimator of p_s. For all $s \in \mathcal{S}$ let $X^s \sim \mathbb{P}_{X|S=s}$ be independent from everything, for all $s', s \in \mathcal{S}$ set the shorthand notation

$$\mathfrak{a}_{ss'} = \mathbf{E}\left|Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(X^s,s)\right)\right) - \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right)\right| \ .$$

Notice that

$$\mathtt{a}_{ss'} = \mathbf{E}\left[\left|Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(X,S)\right)\right) - \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X,S)+\varepsilon\right)\right)\right| |S=s\right] \ ,$$

and therefore we can write

$$\mathbf{E}\left|\hat{g}(X,S) - g^*(X,S)\right| = \mathbf{E}[\Delta(\hat{g})] \leq \sum_{s,s'\in\mathcal{S}} p_s p_{s'} \mathtt{a}_{ss'} + \frac{1}{2}\sum_{s\in\mathcal{S}} \underline{\lambda}_s^{-1} \mathbf{E}\left|p_s - \hat{p}_s\right| \ .$$

Notice that the term $\mathbf{E}\left|p_s - \hat{p}_s\right| = N^{-1}\mathbf{E}|Np_s - V|$, where V is the binomial random variable with parameters (N, p_s), thus using the Cauchy–Schwarz inequality we can write $\mathbf{E}\left|p_s - \hat{p}_s\right| \leq N^{-1}\sqrt{\mathrm{Var}(V)} = \sqrt{p_s(1-p_s)/N}$ and the above bound reads as

$$\begin{aligned}\mathbf{E}\left|\hat{g}(X,S) - g^*(X,S)\right| &\leq \sum_{s,s'\in\mathcal{S}} p_s p_{s'} \mathtt{a}_{ss'} + \frac{1}{2}\sum_{s\in\mathcal{S}} \underline{\lambda}_s^{-1}\sqrt{\frac{p_s(1-p_s)}{N}} \\ &\leq \sum_{s,s'\in\mathcal{S}} p_s p_{s'} \mathtt{a}_{ss'} + \frac{N^{-1/2}}{2}\max_{s\in\mathcal{S}}\underline{\lambda}_s^{-1}\sum_{s\in\mathcal{S}}\sqrt{p_s(1-p_s)} \ . \end{aligned} \tag{15}$$

It remains to bound $\mathtt{a}_{ss'}$ for each $s, s' \in \mathcal{S}$. Fix some $s, s' \in \mathcal{S}$ (they can be equal), then

$$\begin{aligned}\mathtt{a}_{ss'} \leq & \underbrace{\mathbf{E}\left|\hat{Q}_{f^*|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right) - \hat{Q}_{\hat{f}|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right)\right|}_{\mathtt{a}^1_{ss'}} \\ & + \underbrace{\mathbf{E}\left|Q_{f^*|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right) - \hat{Q}_{f^*|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right)\right|}_{\mathtt{a}^2_{ss'}} \\ & + \underbrace{\mathbf{E}\left|Q_{f^*|s'}\left(F_{f^*|s}\left(f^*(X^s,s)\right)\right) - Q_{f^*|s'}\left(\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right)\right|}_{\mathtt{a}^3_{ss'}} \ . \end{aligned} \tag{16}$$

We bound each of the three terms separately.

First term ($\mathtt{a}^1_{ss'}$): Notice that $\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)$ is distributed uniformly on $\{0, 1/|\mathcal{I}_1^s|, 2/|\mathcal{I}_1^s|, \ldots, 1\}$ conditionally on labeled data $\mathcal{L}$ (see *e.g.,* [40, Lemma 13.1]). Thus, we have

$$\mathtt{a}^1_{ss'} = \frac{1}{|\mathcal{I}_1^s|+1}\sum_{j=0}^{|\mathcal{I}_1^s|}\mathbf{E}\left|\hat{Q}_{f^*|s'}\left(\frac{j}{|\mathcal{I}_1^s|}\right) - \hat{Q}_{\hat{f}|s'}\left(\frac{j}{|\mathcal{I}_1^s|}\right)\right| \ . \tag{17}$$

Notice that for all $j \in \{1, \ldots, |\mathcal{I}_1^s|\}$ and all $\alpha \in ((j-1)/|\mathcal{I}_1^s|, j/|\mathcal{I}_1^s|]$ it holds that

$$\hat{Q}_{f^*|s'}\left(\frac{j}{|\mathcal{I}_1^s|}\right) = \hat{Q}_{f^*|s'}(\alpha) \ .$$

The above implies that

$$\frac{1}{|\mathcal{I}_1^s|}\hat{Q}_{f^*|s'}\left(\frac{j}{|\mathcal{I}_1^s|}\right) = \int_{j/|\mathcal{I}_1^s|}^{(j+1)/|\mathcal{I}_1^s|} \hat{Q}_{f^*|s'}(\alpha)\, d\alpha \ , \tag{18}$$

and the same argument repeated for $\hat{Q}_{\hat{f}|s'}$ implies that

$$\frac{1}{|\mathcal{I}_1^s|}\hat{Q}_{\hat{f}|s'}\left(\frac{j}{|\mathcal{I}_1^s|}\right) = \int_{j/|\mathcal{I}_1^s|}^{(j+1)/|\mathcal{I}_1^s|} \hat{Q}_{\hat{f}|s'}(\alpha)\, d\alpha \ . \tag{19}$$

Substituting Eqs. (18)-(19) in Eq. (17) and using Definition C.2 we get

$$\mathtt{a}^1_{ss'} \leq 2\mathbf{E}\int_0^1 \left|\hat{Q}_{f^*|s'}(\alpha) - \hat{Q}_{\hat{f}|s'}(\alpha)\right| d\alpha = 2\mathbf{E}\mathcal{W}_1(\hat{\nu}^0_{f^*|s'}, \hat{\nu}^0_{\hat{f}|s'}) \ ,$$

where for $j=0$ in Eq. (17) we used the fact that $\frac{1}{|\mathcal{I}_1^s|}\mathbf{E}|\hat{Q}_{f^*|s'}(0)-\hat{Q}_{\hat{f}|s'}(0)| \leq \mathbf{E}\int_0^1|\hat{Q}_{f^*|s'}(\alpha)-\hat{Q}_{\hat{f}|s'}(\alpha)|d\alpha$. Using the coupling definition of the Wasserstein distance and the way we have defined $\hat{\nu}^0_{f|s'}$, we can write

$$\mathcal{W}_1(\hat{\nu}^0_{f^*|s'},\hat{\nu}^0_{\hat{f}|s'}) \leq \frac{1}{|\mathcal{I}_0^{s'}|}\sum_{i\in\mathcal{I}_0^{s'}}\left|f^*(X_i^{s'},s')+\varepsilon_{is'}-(\hat{f}(X_i^{s'},s')+\varepsilon_{is'})\right| \quad ,$$

almost surely. Since $\{X_i^{s'}\}_{i\in\mathcal{I}_0^{s'}}$ are i.i.d. from $\mathbb{P}_{X|S=s'}$, then conditionally on $\mathcal{L}$ the random variables $\{|f^*(X_i^{s'},s)-\hat{f}(X_i^{s'},s')|\}_{i\in\mathcal{I}_0^{s'}}$ are i.i.d. . Furthermore, using Lemma C.1 we can write

$$\mathtt{a}^1_{ss'} \leq 2\mathbf{E}\mathcal{W}_1(\hat{\nu}^0_{f^*|s'},\hat{\nu}^0_{\hat{f}|s'}) \leq 2\mathbf{E}\left[\left|f^*(X,S)-\hat{f}(X,S)\right|\big|S=s'\right] \overset{\text{Lemma C.1}}{\leq} 2\mathtt{A}b_n^{-1/2} \quad . \tag{20}$$

Second term ($\mathtt{a}^2_{ss'}$): Note that under Assumption 4.2, the Lipschitz constant of $Q_{f^*|s'}$ is upper bounded by $\underline{\lambda}_{s'}^{-1}$. Then, taking supremum and using Definition C.2 we apply Theorem C.4 to get

$$\mathtt{a}^2_{ss'} \leq \mathbf{E}\mathcal{W}_\infty\left(\nu_{f^*|s'},\hat{\nu}^0_{f^*|s'}\right) \leq a\underline{\lambda}_{s'}^{-1}N_{s'}^{-1/2} \quad . \tag{21}$$

where a is an absolute positive constant ($a=2\sqrt{2\pi}$ is sufficient).

Third term ($\mathtt{a}^3_{ss'}$): We can write, using Assumption 4.2 that

$$\begin{aligned}\mathtt{a}^3_{ss'} \leq& \underline{\lambda}_{s'}^{-1}\mathbf{E}\left|F_{f^*|s}\left(f^*(X^s,s)\right)-\hat{F}_{\hat{f}|s}\left(\hat{f}(X^s,s)+\varepsilon\right)\right| \\ \leq& \underline{\lambda}_{s'}^{-1}\mathbf{E}\left|F_{f^*|s}\left(f^*(X^s,s)\right)-F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right)\right| + \underline{\lambda}_{s'}^{-1}\mathbf{E}\|F_{\bar{\nu}_{\hat{f}|s}}(t)-\hat{F}_{\hat{f}|s}(t)\|_\infty \quad ,\end{aligned} \tag{22}$$

with $F_{\bar{\nu}_{\hat{f}|s}}$ defined for all $s\in\mathcal{S}$ and all $t\in\mathbb{R}$ as

$$F_{\bar{\nu}_{\hat{f}|s}}(t) = \mathbf{P}\left(\hat{f}(X^s,s)+\varepsilon\leq t\big|\mathcal{L}\right) \quad . \tag{23}$$

The second term in Eq. (22) is bounded by $\lesssim 2\underline{\lambda}_{s'}^{-1}N_s^{-1/2}$ thanks to the Dvoretzky–Kiefer–Wolfowitz inequality recalled in Theorem B.1. Thus, it remains to bound the first term in Eq. (22). We introduce the following shorthand notation for the first term in Eq. (22)

$$(*) = \mathbf{E}\left|F_{f^*|s}\left(f^*(X^s,s)\right)-F_{\bar{\nu}_{\hat{f}|s}}\left(\hat{f}(X^s,s)+\varepsilon\right)\right| \quad .$$

Let $\tilde{X}^s\sim\mathbb{P}_{X|S=s}$ and $\tilde{\varepsilon}\sim U[-\sigma,\sigma]$ be independent from $\varepsilon, X^s, \mathcal{L}$ and each other. Based on this notation we can write

$$(*)=\mathbf{E}\Big|\underbrace{\mathbf{P}\left(f^*(\tilde{X}^s,s)-f^*(X^s,s)\leq 0\big|\varepsilon,X^s,\mathcal{L}\right)}_{H_0}-\underbrace{\mathbf{P}\left(\hat{f}(\tilde{X}^s,s)+\tilde{\varepsilon}\leq\hat{f}(X^s,s)+\varepsilon\big|\varepsilon,X^s,\mathcal{L}\right)}_{H_1}\Big| \quad . \tag{24}$$

Furthermore, if $\Delta(X^s)=f^*(X^s,s)-\hat{f}(X^s,s)$, $\Delta(\tilde{X}^s)=f^*(\tilde{X}^s,s)-\hat{f}(\tilde{X}^s,s)$, and $\Delta_\varepsilon=\varepsilon-\tilde{\varepsilon}$, then simple algebra yields

$$H_1 = \mathbf{P}\left(f^*(\tilde{X}^s,s)-f^*(X^s,s)\leq\Delta_\varepsilon+\Delta(\tilde{X}^s)-\Delta(X^s)\big|\varepsilon,X^s,\mathcal{L}\right) \quad .$$

For all $a,b\in\mathbb{R}$ it holds that $|\mathbf{1}_{\{a\leq 0\}}-\mathbf{1}_{\{a\leq b\}}| \leq \mathbf{1}_{\{0\wedge b\leq a\leq 0\vee b\}} \leq \mathbf{1}_{\{-|b|\leq a\leq|b|\}} = \mathbf{1}_{\{|a|\leq|b|\}}$. Applying this fact to Eq. (24) with $a=f^*(\tilde{X}^s,s)-f^*(X^s,s)$ and $b=\Delta_\varepsilon+\Delta(\tilde{X}^s)-\Delta(X^s)$ we get

$$\begin{aligned}(*) \leq& \mathbf{P}\left(\left|f^*(\tilde{X}^s,s)-f^*(X^s,s)\right|\leq|\Delta_\varepsilon|+|\Delta(\tilde{X}^s)|+|\Delta(X^s)|\right) \\ \leq& \mathbf{P}\left(\left|f^*(\tilde{X}^s,s)-f^*(X^s,s)\right|\leq 3|\Delta_\varepsilon|\right)+\mathbf{P}\left(\left|f^*(\tilde{X}^s,s)-f^*(X^s,s)\right|\leq 3|\Delta(\tilde{X}^s)|\right) \\ &+\mathbf{P}\left(\left|f^*(\tilde{X}^s,s)-f^*(X^s,s)\right|\leq 3|\Delta(X^s)|\right) \quad .\end{aligned}$$

By definition of $\tilde{X}^s$ the random variables $X^s, \tilde{X}^s$ are exchangeable, hence

$$\mathbf{P}(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 3|\Delta(\tilde{X}^s)|) = \mathbf{P}(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 3|\Delta(X^s)|) \ .$$

Furthermore, using the fact that $|\varepsilon - \tilde{\varepsilon}| \leq 2\sigma$ almost surely we get

$$(*) \leq \mathbf{P}\left(\left|f^*(\tilde{X}^s, s) - f^*(X^s, s)\right| \leq 6\sigma\right) + 2\mathbf{P}\left(\left|f^*(\tilde{X}^s, s) - f^*(X^s, s)\right| \leq 3\left|\Delta(X^s)\right|\right) \ . \quad (25)$$

Thanks to Assumption 4.2 we have the following bound on the first term in Eq. (25)

$$\mathbf{P}\left(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 6\sigma\right) \leq \mathbf{E}\left[\mathbf{P}\left(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 6\sigma | X^s\right)\right] \leq 12\overline{\lambda}_s\sigma \ .$$

For the second term in Eq. (25), we observe that Assumption 4.2 yields almost surely

$$\mathbf{P}\left(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 3\left|\Delta(X^s)\right| \middle| \mathcal{L}, X^s\right) \leq 6\overline{\lambda}_s\left|\Delta(X^s)\right| \ .$$

Thus, taking the total expectation on both sides of this inequality we get

$$\mathbf{P}\left(|f^*(\tilde{X}^s, s) - f^*(X^s, s)| \leq 3\left|\Delta(X^s)\right|\right) \leq 6\overline{\lambda}_s\mathbf{E}\left|\Delta(X^s)\right| \overset{\text{Lemma C.1}}{\leq} 6\overline{\lambda}_s\mathbb{A}b_n^{-1/2} \ .$$

Since $\sigma \lesssim b_n^{-1/2}$, then we have demonstrated that $(*) \lesssim \overline{\lambda}_s b_n^{-1/2}$. Substituting this bound into Eq. (22), we derive that

$$\mathfrak{a}_{ss'}^3 \lesssim \underline{\lambda}_{s'}^{-1}\overline{\lambda}_s b_n^{-1/2} + \underline{\lambda}_{s'}^{-1} N_s^{-1/2} \ . \quad (26)$$

Gathering three terms together: Finally, substituting Eqs. (20), (21), (26) into Eq. (16) we get

$$\mathfrak{a}_{ss'} \lesssim b_n^{-1/2} + \underline{\lambda}_{s'}^{-1}\overline{\lambda}_s b_n^{-1/2} + \underline{\lambda}_{s'}^{-1} N_{s'}^{-1/2} + \underline{\lambda}_{s'}^{-1} N_s^{-1/2} \ .$$

Finally, substituting the bound above into Eq. (15) we arrive at

$$\begin{aligned}
\mathbf{E}\left|\hat{g}(X, S) - g^*(X, S)\right| \lesssim & b_n^{-1/2} + \left(\sum_{s\in\mathcal{S}} p_s\underline{\lambda}_s^{-1}\right)\left(\sum_{s\in\mathcal{S}} p_s\overline{\lambda}_s\right) b_n^{-1/2} \\
& + \sum_{s\in\mathcal{S}} p_s\underline{\lambda}_s^{-1} N_s^{-1/2} + \left(\sum_{s\in\mathcal{S}} p_s\underline{\lambda}_s^{-1}\right)\left(\sum_{s\in\mathcal{S}} p_s N_s^{-1/2}\right) \\
& + N^{-1/2}\max_{s\in\mathcal{S}}\underline{\lambda}_s^{-1}\sum_{s\in\mathcal{S}}\sqrt{p_s(1-p_s)} \\
\lesssim & b_n^{-1/2} + \sum_{s\in\mathcal{S}} p_s N_s^{-1/2} + \sqrt{|\mathcal{S}|}N^{-1/2} \ ,
\end{aligned}$$

where in the last inequality we used the fact that

$$\sum_{s\in\mathcal{S}}\sqrt{p_s(1-p_s)} \leq \sum_{s\in\mathcal{S}}\sqrt{p_s} \leq \sqrt{|\mathcal{S}|}\sqrt{\sum_{s\in\mathcal{S}} p_s} = \sqrt{|\mathcal{S}|} \ .$$

This ends the proof. □

Remark C.5. *Notice that the exact constant in front of the rate of convergence in Theorem 4.4 can be recovered following the proof. Furthermore, this proof can be extended to control L_p norm*

$$\left(\mathbf{E}|g^*(X, S) - \hat{g}(X, S)|^p\right)^{1/p} \ ,$$

for all $p \in [1, \infty)$ (the current proof deals only with $p = 1$). To achieve it one only needs to extend Lemma C.1 while the rest of the proof follows line-by-line using deviation results on Wasserstein-p distance on the real line [7]. Finally, it is possible to extend this result under the same assumptions to control $\mathbf{E}\left\|g^* - \hat{g}\right\|_\infty$, which induces an extra multiplicative polylogarithmic factor in $b_n^{-1/2}$.